

DOCUMENT RESUME

ED 184 901

SE 030 639

AUTHOR Mandaville, Garrett K.
TITLE Teaching Statistics Using SAS.
PUB DATE Apr 80
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 7-11, 1980). Contains marginal legibility in Charts and Tables.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Mathematics; *Computer Oriented Programs; *Computer Programs; *Educational Research; Higher Education; *Mathematical Applications; Mathematics Curriculum; Mathematics Instruction; Research; Statistical Analysis; *Statistics
IDENTIFIERS *Mathematics Education Research

ABSTRACT

The Statistical Analysis System (SAS) is presented as the single most appropriate statistical package to use as an aid in teaching statistics. A brief review of literature in which SAS is compared to SPSS, BMDP, and other packages is followed by six examples which demonstrate features unique to SAS which have pedagogical utility. Of particular merit is the capability to create simulated data sets which may be used to: (1) demonstrate characteristics of sampling distributions, (2) provide individualized data sets for student use, and (3) allow for discussing of Type 1 errors and power. In the final example, the MATRIX procedure is used to provide various solutions to the single-factor repeated measures design. (Author/MK)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED184901

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Garrett Mandeville

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

TEACHING STATISTICS USING SAS*

Garrett K. Mandeville
University of South Carolina

BEST COPY AVAILABLE

* One of five presentations in the symposium "Using the Computer in Teaching Statistics" given at the annual convention of the American Educational Research Association in Boston, Ma., April 7-11, 1980.

SE 030 639

Summary

The statistical Analysis System (SAS) is presented as the single most appropriate statistical package to use as an aid in teaching statistics. A brief review of literature in which SAS is compared to SPSS, BMDP, and other packages is followed by six examples which demonstrate features unique to SAS which have pedagogical utility. Of particular merit is the capability to create simulated data sets which may be used to (1) demonstrate characteristics of sampling distributions, (2) provide individualized data sets for student use and (3) allow for discussions of Type 1 error and power. In the final example the MATRIX procedure is used to provide various solutions to the single factor repeated measures design.

Introduction and Objectives

The teaching of applied statistics at either an introductory or an advanced level is facilitated by computational tools such as calculators and computers. Although I haven't researched it, I suspect that the vast majority of applied statistics courses taught at colleges and universities where a computer is reasonably accessible, have some form of activity requiring the student to confront the computer for purposes of data analysis. At those institutions which have made a large investment in CAI, the reasonable solution is probably to use some sort of interactive system; at others, the instructor has a choice of using one or more of the available software packages for statistical computing and/or developing his or her own software package and using "batch processing".

This paper will be instructive to those individuals teaching statistics at institutions of the second type described above in that it describes many of the ways that the flexibility of the current version of the Statistical Analysis System (Helwig and Council, 1979) lead to it's being selected as the logical choice if a single "package" is to be considered. The writer also recommends the "single package" approach, particularly in the introductory courses, so that students need to learn only one syntactical system.

Background

In teaching a course in statistics there are various types of tasks which the instructor might wish to relegate to the computer and various computational capabilities which should be reasonably expected of a computer package. For purposes of discussion, I will consider the following two rather

broad categories: (1) data simulation and (2) computational and data manipulative flexibility.

There are at least two situations in which the capability to simulate data would be useful. First, given the importance of the concept of a sampling distribution, it is useful to confront student's with "in hand" evidence of the validity of truths such as those couched in the Central Limit Theorem. On the other hand, for purposes of generation of data sets for classroom discussion and/or homework problems, the instructor can maintain control of the parameters of generation and each student can have his/her own individualized printout. During classroom discussion sessions, notions of Type I and Type II error and power can be dealt with on an empirical basis.

Regarding computational/manipulative flexibility, we should expect the capacity to (1) calculate somewhat "unusual" statistics, (2) to test assumptions of various statistical procedures, (3) to perform analyses on easily created partitions of the data set, (4) to pass data sets from step-to-step if a sequential analytic strategy is necessary, and (5) to merge data sets gathered at different times (such as in longitudinal studies). Although not exhaustive, the above are reasonable minimal expectations.

Recently many researchers have compared various of the packages regarding their adequacy on dimensions such as ease of use, accuracy and flexibility. These comparisons include computer graphics (Ryan and Vellman, 1977), regression analysis in terms of ease (Allen and Vellman, 1977) and accuracy (Vellman, Seaman and Allen, 1977), completeness of ANCOVA output in mixed models (Heiberger, 1976; Federer and Henderson, 1978) and multivariate analysis of variance (Hohwald and Heiberger, 1977). The most common packages involved in those comparisons were SAS (Barr, et al., 1976), BMDP (Dixon, 1975) and

SPSS (Nie, et al, 1975). P-STAT (Buhler, 1976) and MINITAB (Ryan, et al, 1976) were also quite popular in these studies. Since these studies were performed prior to the release of SAS79, various version of SAS76 were involved. If one compares the performance of the three most common packages (SAS, BMDP, SPSS), the results tend to support the position taken in this paper. With respect to graphics, the old SAS PROC SCATTER satisfied the 13 standards for plots (since the 1979 CHART procedure was not available, SAS was not evaluated with respect to histograms). For general "handiness" the technical, non-tutorial nature of the SAS manual (slightly improved with SAS79) was noted but the quality of references, clarity of error messages and ability to interface with other systems were viewed as strong points. When presented with ill-conditioned data in a regression problem, SAS produced the most informative messages to the user and was as accurate as BMDP9R (the other packages were less accurate). For the mixed model ANCOVA comparison, GENSTAT was viewed as providing the most useful output but SAS provided output for more features than did BMDP or SPSS. (It is likely that enhancements in SAS GLM in the 1979 version might have changed these results.) The final study cited above dealt with MANOVA and no SPSS routine was available for consideration. SAS GLM was seen to be about as complete as MULTIVARIANCE and faster. The computing routines used fewer CPU seconds.

Even though an outdated version of SAS was used in the above studies, this writer believes that a synthesis of the results would suggest SAS if a decision as to which one of the packages to adopt were to be made. The above studies, however, either dealt with "standards" for packages, or accuracy or flexibility in output for standard statistical problems. In the remainder of this paper, the writer, capitalizing on many new features of SAS79, will present examples of the utility of SAS beyond the situations presented in the above studies.

Examples

The following section provides a list of examples of how the SAS system can be used for a variety of tasks of the type suggested above. In each case below, a brief introduction to the problem is provided. A listing of the SAS statements used is in Appendix 1 and example output is provided in displays at the end of the paper. In the former SAS comment statements, which begin with an asterisk (*) are used for purposes of documentation. Although the SAS log is omitted, total CPU time for each example using release 79.2B on an Amdahl 470/VS computer is provided. Although sets of SAS statements are not strictly "programs," they will be referred to as such throughout the descriptions.

Examples of Data Simulation

Example 1. Normal approximation to the binomial distribution (See Display 1)

Using the binomial parameters N and P and the number of samples (NSAMP), this program generates NSAMP samples from the prescribed binomial distribution. Output includes a simultaneous plot of the empirical relative frequency distribution and the theoretical distribution based on the normal approximation. The program would be useful in an introductory statistics class to demonstrate the conditions on N and P for which the normal approximation is reasonable.

Example 2: Comparison of empirical and theoretical t distributions

This program requires that the sample size (N) and the number of samples (NSAMP) are input. NSAMP samples of size N are simulated from the unit normal distribution and the t statistic for testing $H_0: \mu = 0$ is calculated for each. The values of the t statistic are cast into a frequency distribution with a class interval width arbitrarily chosen to be 0.4 units. A simultaneous plot of the empirical and theoretical frequency distributions is output. The results could be used to indicate the validity of the t distribution and, although currently not programmed, the platykurtic nature of the t distribution for small degrees of freedom.

Example 3: Simulation of RSD of r and Fisher's Z_r

This program requires the number of (X,Y) pairs, i.e., the sample size (N), the population correlation between X and Y (ρ), and the number of data sets to be generated (NSAMP). NSAMP bivariate, normal data sets of size N with mean 0 and standard deviation 1 on X and Y and a correlation of ρ between X and Y are created. As output two plots are included; (1) the (generally asymmetric) empirical sampling distribution of the Pearson r and, (2) the approximately normal distribution of the correlations when transformed via Fisher's Z_r transformation. In the latter case the appropriate normal distribution is overprinted for comparison purposes. The program includes a locally produced RECODE procedure and is geared toward values of ρ in the .6 to .8 range.

Example 4: Simulation and output of "individualized" data sets

The following example is prepared to demonstrate how SAS can be used to output simulated data sets for each designated student so as to (1) provide individualized student data sets for purposes of computing practice and, (2) facilitate discussion of the power of a statistical test. Although the example is from the rather exciting single sample normal or t test, the approach of outputting individual data sets could be, without difficulty, incorporated into the earlier examples or other situations the instructor might wish to consider. The program requires as input (1) the names (or some other ID such as SS numbers) of the students for which output is to be generated and (2) the number of samples to be generated (NSAMP=number of students identified above), the true mean (MUT), the hypothesized mean (MUH) and the (known) population standard deviation (SIGMA). The sample size in the coding presented is set at 20. For each student the sample data

set, the sample mean and standard deviation, the values of the Z statistic (assuming a known standard deviation), it's one-tailed p-value and the t statistic and it's p-value are reported. The program does not output results aggregated across students; that's the instructor's job. Although the one sample t-test is probably the more practical of the two procedures, the normal test is also included because of the relative simplicity of power calculations.

Examples of Data Manipulation/Computation

Although many of the earlier programs have utilized aspects of the SAS data manipulative capacity, particularly regarding the creation of data sets, they might have escaped the reader. Therefore, the following specific examples, not related to data simulation and therefore possibly more instructive, are provided. The data sets utilized have been kept small for purposes of space.

Example 5: Scoring and outputting classical item statistics

Although most researchers involved in item analysis activities probably use a program they have selected or created to suit their own personal needs, it is worthwhile to know that SAS can be easily programmed to "score" a test and output the common classical item and test statistics. The program included is written to score a ten item multiple choice test, to output student raw scores, standard scores (mean 50, standard deviation 10) and percentile rank scores. In addition using the standard SAS PROC CORR it provides item difficulties (proportion correct), discriminations (point biserial with total score), and the raw score mean and standard deviation. Finally the KR-20 reliability of the test is calculated and output. An alphameric ID is assumed on the student records and the test key precedes the first student

record. Other useful summaries such as a histogram of the score distribution could be included if desired.

Example 6: McNemar's test for change in correlated proportions

Although the following example is artificial, the motivation to write the program was very practical. Given pretest and posttest data on a test item, it is reasonable to expect that the proportion correct should increase from pretest to posttest if the instruction during the interval between the two testings was relevant to the objective the item was measuring. (This, of course, assumes that the item was not so easy that a large majority of the students answered the item correctly on the pretest.) Information on the change in the pretest to posttest proportion correct (and the significance of the change) would then be useful to identify either (1) items which, possibly due to some flaw in construction, did not change as a result of instruction or (2) instructional objectives which were taught poorly or not taught at all.

The program assumes pretest and posttest data for a five item test and uses the input and scoring strategy from example 5 above for both pretest and posttest. McNemar's test for change in correlated proportions is the only item statistic reported.

Example 7: Single factor repeated measures design

The program first provides summary statistics (mean vector and covariance matrix) and the results for the univariate F test of equality of the elements of the mean vector following the test statistic for the Mauchly test of sphericity (an assumption of the univariate ANOVA). Next, two alternative estimates of epsilon, the quantity by which the numerator and denominator degrees of freedom of the reference F distribution for the univariate F statistic could

be adjusted and the adjusted tests corresponding to these two estimates are provided. Finally, the multivariate F statistic based on Hotelling's T^2 is output. Thus, a rather complete picture of the results of competing procedures is provided. The program as written assumes three levels for the repeated measure.

The examples dealing with computation and data manipulation may not have created in the reader the full appreciation of the flexibility of SAS. One feature of particular utility which may have gone unnoticed is the feature that any SAS procedure can output results for all levels or combinations of levels of one or more independent variables. This is facilitated by sorting the data set on the relevant variable(s) and then including the corresponding variable(s) in a BY statement accompanying the description of the procedure. This is useful, for example, if a researcher wishes to compare the correlation between two variables for two or more independent groups.

A second very useful feature of SAS is the creation of output data sets from various procedures. In the examples above this feature of PROC SUMMARY and PROC MEANS has been capitalized on. A third feature of SAS is the ease with which data sets created at different times may be merged and the corresponding creation of output data sets which satisfy different conditions with respect to the completeness of the data. For example, if the researcher wished to, he or she could create a data set for which pretest but no posttest records were available, one for which posttest but no pretest records were available, and a "clean" data set for which both pretest and posttest records were available. These might be very useful in the assessment of the external validity of the research.

Finally it should be noted that PROC MATRIX provides, for the more sophisticated user, computational power which should be sufficient to tackle all but the most complex data analytic situations.

Summary

Although surely not exhaustive, the above examples have been chosen in an attempt to highlight the flexibility of the SAS system as a tool to aid in statistics instruction and statistical computation. Surely other systems provide procedures to perform some of the above functions. For example, McNemar's Test is available in SPSS and various repeated measures designs can be analyzed using BMD. Neither of these systems, however, has the data simulation capability necessary to provide the type of output described earlier. Many years ago the writer prepared an extensive FORTRAN program to simulate data to distribute in class; he only wishes that SAS79 had been available at that time.

With respect to syntax, an objective observer would probably conclude that the syntax of SAS is about as straightforward as possible, surely easier to "pick up" than the syntax of SPSS or BMD. The handiness of the SAS log being separate from the output file also is a useful feature, especially for the researcher who accesses his output on a low speed terminal and can, based on this information, make a decision whether the output is worth routing to a high speed printer.

Finally it has been my experience that most researchers (this includes students, colleagues, etc.), given exposure to various packages, select one, learn it quite well, and use it for most of their data analytic problems. At some point, they drop whatever passing interest they might have had in the development of other systems and the die is cast. Although this is reasonable behavior, I think it is unfortunate.

It is my belief that the majority of us involved in teaching statistics have probably been exposed to SPSS either in our training or subsequent to it; fewer will have been exposed to BMD and fewer yet are likely to have retained an interest in it. Those of us who were also exposed to SAS had to deal with

extremely terse manuals describing the 1972 and 1976 versions which had virtually no pedagogical intent. Both the original and current SPSS Manuals, however, are quite readable and have included information useful to the student. For these reasons, and the wider availability of SPSS in the earlier years, I expect that the majority of us who have the responsibility of teaching applied statistics view SPSS as "our" package and employ the system in our classes.

Although the current SAS manual is much better organized and much more comprehensible than the earlier ones, it contains less in the way of rationale or examples relevant to social science research than the SPSS manual. On the other hand history suggests that those responsible for SAS are more likely to make regular additions and improvements to the system than is true for SPSS (both SAS76 and SAS79 postdate the SPSS Manual of 1975). Although new versions of SPSS are upgraded, the documentation is handled in updates which are difficult to obtain and confusing and an extra financial burden on students.

Additionally the SAS support staff includes quality statisticians who are, probably, more aware of the current status of the applied literature. Some of the new procedures in SAS79, such as CHART, SUMMARY and FUNCAT attest to my perception of the SAS philosophy to include the most current and relevant statistical methods. Finally, anyone who has had to deal with large amounts of data will appreciate the difference in CPU time for the same problem under the two systems. SAS, primarily coded PL/I and assembler, is much faster than the FORTRAN based SPSS for processing large data sets.

In this paper I have attempted to present some examples which highlight the flexibility of the SAS system with particular emphasis on classroom instruction. In the discussion in the summary above I have compared SAS and SPSS. Although other systems could have been included, they tend to be less

general in scope and less widely available and should provide no serious competition to SAS or SPSS as a package to employ in the early training of educational researchers.

I have no axe to grind with SPSS; nor do I have anything to gain by promoting the SAS system. I simply believe it is the better of the two systems and that, if we desire to provide the best training for our students, it should be an integral part of our instruction.

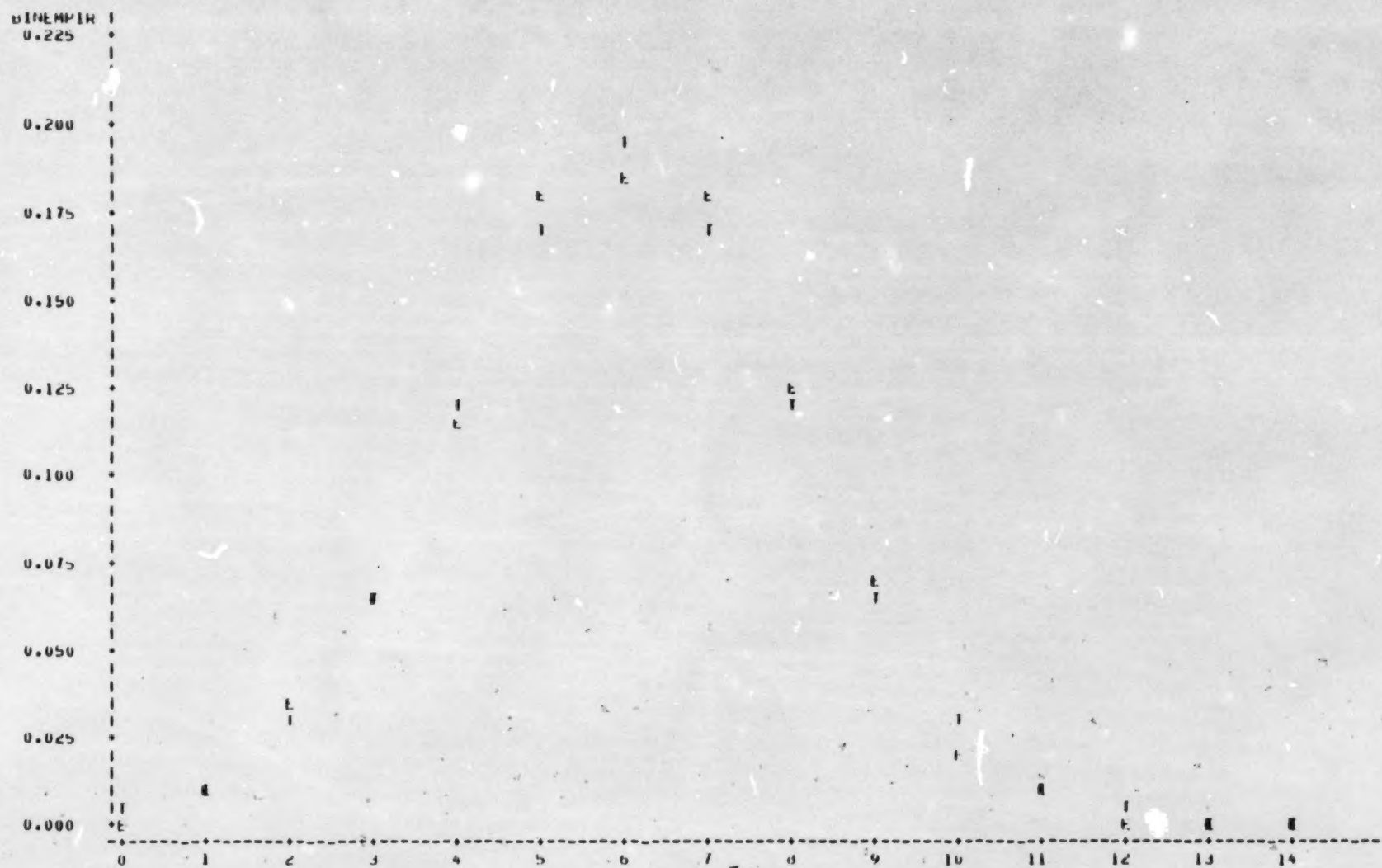
References

- Allen, I. E. and Vellman, P. F. "The Handiness of Package Regression Routines," Proceedings Statistical Computing Section, American Statistical Association, 1977, 95-101.
- Barr, A. J., Goodnight, J. H., Sall, J. P. and Helwig J. T. A User's Guide to SAS 76, Raleigh, N.C.: SAS Institute, 1976.
- Buhler, R. and Buhler, S. P-STAT - A Computing System for File Manipulation and Statistical Analysis, Version 3.07, 1976.
- Dixon, W. J. (ed.), BMDP Biomedical Computer Programs, Berkeley: University of California Press, 1975.
- Federer, W. T. and Henderson, H. V. "Covariance Analysis of Designed Experiments Using Statistical Packages," Proceedings Statistical Computing Section, American Statistical Association, 1978, 332-337.
- Finn, J. D. MULTIVARIANCE: Univariate and Multivariate Analysis of Variance and Covariance, and Regression, Chicago: National Educational Resources, Inc., 1976.
- Heiberger, R. M. "Conceptualization of Experimental Designs and their Specification and Computation with ANOVA Programs," Proceeding, Statistical Computing Section, American Statistical Association, 1976, 13-23.
- Helwig, J. T. and Council, K. A. (eds.), SAS User's Guide, Raleigh, N.C: SAS Institute, 1979
- Hohwald, J. and Heiberger, R. M. "An Evaluation of Manova Programs," Proceedings Statistical Computing Section, American Statistical Association, 1977, 250-255.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K. and Bent, D. H., Statistical Package for the Social Sciences, New York: McGraw-Hill Publishing Company, 1975.
- Ryan, B. F. and Vellman, P. "Computer Graphics in Statistical Software," Proceedings, Statistical Computings Section, American Statistical Association, 1977, 48-55.
- Ryan, T., Joiner, B., and Ryan, B., Minitab II Reference Manual, Statistics Department, Pennsylvania State University, 1976.
- Vellman, P. F., Seaman J., and Allen I. A. "Evaluating Package Regression Routines," Proceedings Statistical Computing Section, American Statistical Association, 1977, 82-91.

STATISTICAL ANALYSIS SYSTEM

PLOT OF BINEMPIR#1
PLOT OF NUMMP#1

SYMBOL USED IS E
SYMBOL USED IS T



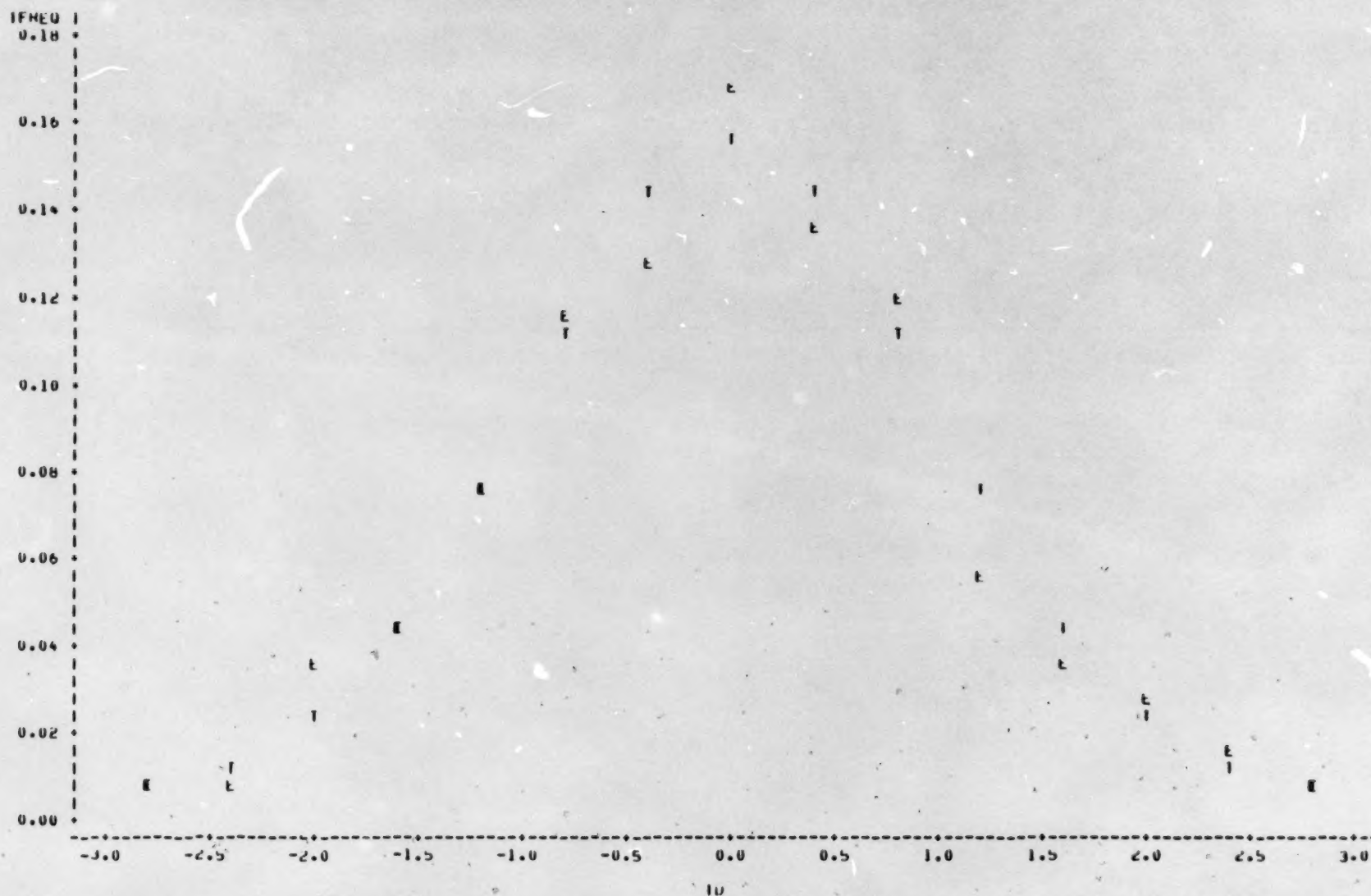
-13-

Display 1: Plot of empirical(E) RSD of binomial N=30, p=.3 Based on NSAMP=500 samples with the normal approximation to the binomial(T) overprinted. Time=1.02 CPU sec.

STATISTICAL ANALYSIS SYSTEM

PLOT OF IFREQ*TD
PLOT OF IFREQ*TD

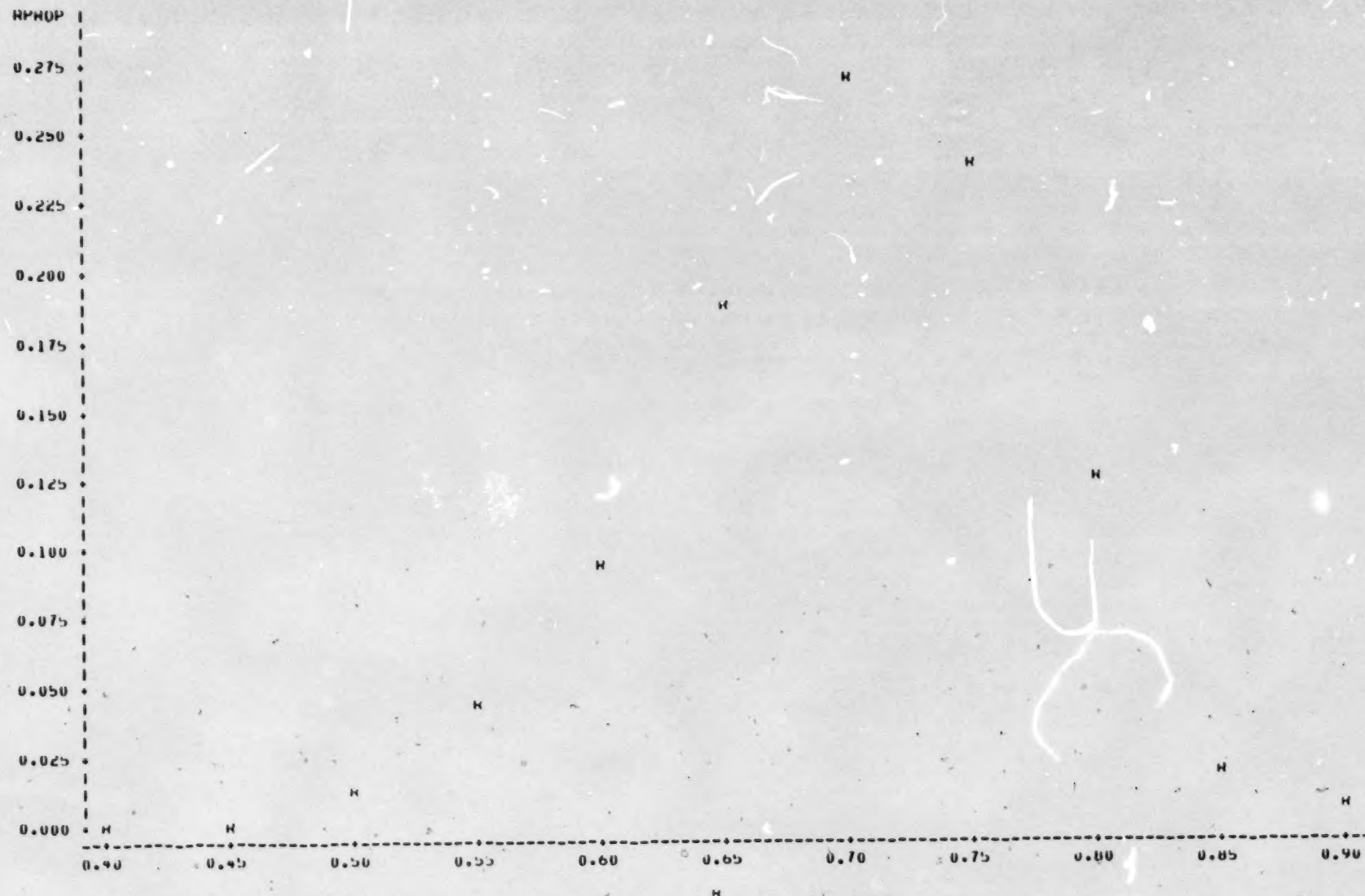
SYMBOL USED IS E
SYMBOL USED IS T



Display 2: Plot of empirical(E) RSD of t statistic for single sample test of a mean based on NSAMP=500. Samples of size N=20. The theoretical t distribution(T) is overprinted. Time=1.64 sec.

STATISTICAL ANALYSIS SYSTEM

PLOT OF HPROP*H SYMBOL USED IS H

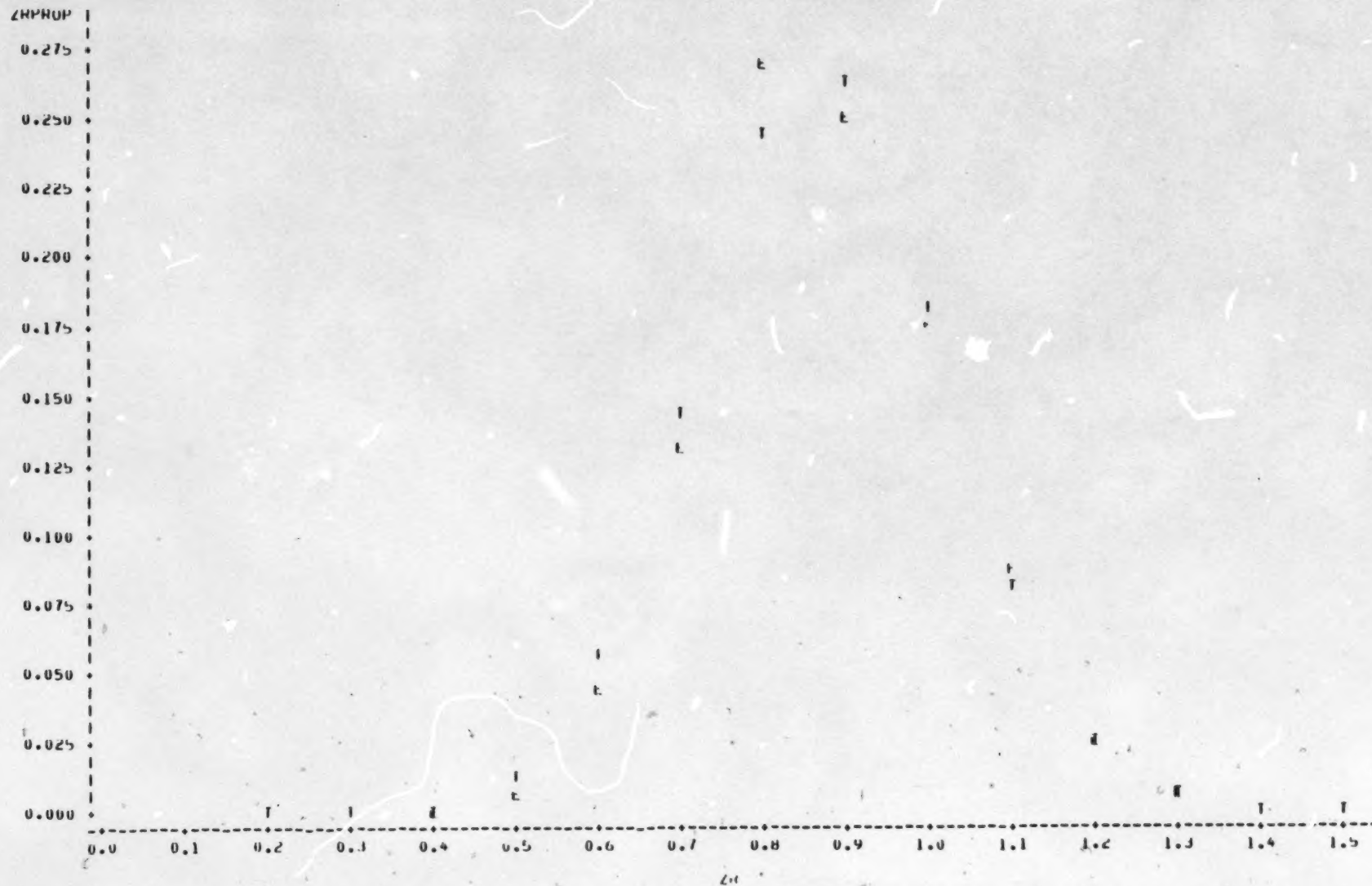


Display 3A: Empirical RSD of r for NSAMP=500. Samples of size 50 from (X,Y) = BIVN (0, 0, 1, 1, RHO = .7).
Time for example 3, 5.48 sec.

STATISTICAL ANALYSIS SYSTEM

PLOT OF ZHPROP*ZH
 PLOT OF NUMMP*ZH

SYMBOL USED IS E
 SYMBOL USED IS T



Display 3B: Empirical RSD of $Z_r(E)$ and theoretical (fitting) normal distribution(T) for data described in Display 3A.

STATISTICAL ANALYSIS SYSTEM
OUTPUT FOR STUDENT JONES E S

THE FOLLOWING IS A SAMPLE FROM A NORMAL DISTRIBUTION WITH A MEAN= 50 AND SIGMA=10

THE HYPOTHESIZED MEAN WAS 45

44 38 44 60 42 55 51 47 52 43 36 39 56 45 66 50 32 35 50 52

THE SAMPLE MEAN AND STANDARD DEVIATION ARE AS FOLLOWS: $\bar{X}=46.85$ $S=8.756$

THE Z STATISTIC (KNOWN SIGMA) AND IT'S 1-TAILED P-VALUE ARE: $Z=0.83$ $P=0.204$

THE T STATISTIC AND IT'S 1-TAILED P VALUE ARE: $T=0.94$ $P=0.178$

STATISTICAL ANALYSIS SYSTEM
OUTPUT FOR STUDENT SMITH F G

THE FOLLOWING IS A SAMPLE FROM A NORMAL DISTRIBUTION WITH A MEAN= 50 AND SIGMA=10

THE HYPOTHESIZED MEAN WAS 45

48 65 36 59 54 59 52 53 36 50 58 50 33 52 50 40 54 58 57 49

THE SAMPLE MEAN AND STANDARD DEVIATION ARE AS FOLLOWS: $\bar{X}=50.65$ $S=8.555$

THE Z STATISTIC (KNOWN SIGMA) AND IT'S 1-TAILED P-VALUE ARE: $Z=2.53$ $P=0.006$

THE T STATISTIC AND IT'S 1-TAILED P VALUE ARE: $T=2.95$ $P=0.004$

Display 4: Listing of output from example 4 for two students. As noted in the printout, each student is provided with a sample of 20 observations from $N(\mu=50, \sigma=10)$ and the results of one sample t or Z tests of $H_0: \mu=45$. Time for four students was 0.55 sec. (Note: Although the pages have been combined, in practice each student would receive his own page of the printout.)

STATISTICAL ANALYSIS SYSTEM

VARIABLE	N	MEAN	STD DEV	SUM	MINIMUM	MAXIMUM
AD1	10	0.70000000	0.48304589	7.00000000	0	1.00000000
AD2	10	0.50000000	0.52704628	5.00000000	0	1.00000000
AD3	10	0.40000000	0.42163702	4.00000000	0	1.00000000
AD4	10	0.50000000	0.52704628	5.00000000	0	1.00000000
AD5	10	0.50000000	0.52704628	5.00000000	0	1.00000000
AD6	10	0.30000000	0.48304589	3.00000000	0	1.00000000
AD7	10	0.40000000	0.51639778	4.00000000	0	1.00000000
AD8	10	0.50000000	0.52704628	5.00000000	0	1.00000000
AD9	10	0.40000000	0.51639778	4.00000000	0	1.00000000
AD10	10	0.40000000	0.51639778	4.00000000	0	1.00000000
TOT	10	5.00000000	2.94392029	50.00000000	1.00000000	10.00000000

CORRELATION COEFFICIENTS / PROB > |R| UNDER H0:RHO=0 / N = 10

	AD1	AD2	AD3	AD4	AD5	AD6	AD7	AD8	AD9	AD10
TOT	0.00000 1.0000	0.50128 0.1399	0.62660 0.0526	0.78773 0.0068	0.78773 0.0068	0.70321 0.0233	0.43853 0.2049	0.50128 0.1399	0.73088 0.0163	0.73088 0.0163
OBS	ID	TOT	STAN	PCTLNK						
1	AAA	2	39.8095	22						
2	BBB	2	39.8095	22						
3	CCC	6	53.3968	59						
4	DDD	6	53.3968	59						
5	EEE	10	66.9842	90						
6	FFF	8	60.1905	81						
7	GGG	7	56.7937	72						
8	HHH	5	50.0000	45						
9	III	3	43.2063	36						
10	JJJ	1	36.4127	9						
					OBS	VTOT	SUMVAN		KH20	
					1	8.66667	2.55556		0.783476	

Display 5: Test score summary for ten students and test and item statistics for a ten item multiple choice test.
Time = 1.21 sec. Again, the pages of the printout have been combined to reduce reproduction costs.

Obs	R1	R2	R3	R4	R5	W1	W2	W3	W4	W5
1	5	3	5	4	6	0	2	0	2	0

MC1	MC2	MC3	MC4	MC5
2.23607	0.447214	2.23607	0.816497	2.44949

Display 6: Results of McNemar's test of change for correlated proportions based on pretest and posttest data for a five item test for a sample of ten students. Time = 0.86 sec.

STATISTICAL ANALYSIS STATISTICAL ANALYSIS STATISTICAL ANALYSIS

NOHS	COL1		
ROW1	10		
NOAH	COL1		
ROW1	3		
AVE	COL1	COL2	COL3
ROW1	22.5	26.7	19.9
LOV	COL1	COL2	COL3
ROW1	151.389	21.3889	16.3889
ROW2	21.3889	4.01111	2.96667
ROW3	16.3889	2.96667	2.76667
FUV	COL1		
ROW1	3.00795		
NV1	COL1		
ROW1	2		
DFE	COL1		
ROW1	18		
PROH0	COL1		
ROW1	0.0746386		

MAUCHLW	COL1
ROW1	0.00149076
EPSHAT	COL1
ROW1	0.50436
PROHEPSH	COL1
ROW1	0.116885
DFD	COL1
ROW1	9.07848
DFN	COL1
ROW1	1.00872
EPSTIL	COL1
ROW1	0.506001
PROHEPST	COL1
ROW1	0.116885
DFD	COL1
ROW1	9.10802
DFN	COL1
ROW1	1.012

TSQR	COL1
ROW1	685.717
FMV	COL1
ROW1	304.763
PROH0V	COL1
ROW1	2.8167E-08

```
* LISTING OF SAS STATEMENTS FOR EXAMPLE 1;
DATA BIN(KEEP=SUCCESS NSAMP) FITNORM(KEEP=I NORMP);
INPUT N P NSAMP;
DO I=1 TO NSAMP;*GENERATE NSAMP SAMPLES FROM BINOMIAL;
SUCCESS=0;
DO J=1 TO N;
X=0; IF UNIFORM(13579) LE P THEN X=1;
SUCCESS+X; END;
OUTPUT BIN; END;
MU=N*P; SIG=SQRT(N*P*(1-P));*CALC PARAMETERS OF NORMAL DIST;
SUMCUM=0;
DO I=0 TO N;
Z=(I-.5-MU)/SIG;CUM=PROBNORM(Z);*CALC Z VALUES FOR EACH POS. # SUCCESS;
IF 0 LE I LE (N-1) THEN DO;
NORMP=CUM-SUMCUM; SUMCUM=NORMP;
OUTPUT FITNORM; END;
ELSE DO;
NORMP=1-SUMCUM; OUTPUT FITNORM; END;
END;
CARDS;
20 .3 500
PROC SUMMARY DATA=BIN; CLASS SUCCESS; VAR NSAMP;
OUTPUT OUT=BINSUM MEAN=NSAMP1;
DATA TEMP(RENAME=(SUCCESS=I));
SET BINSUM; IF TYPE =0 THEN DELETE;DROP _TYPE_ NSAMP1 _FREQ_;
BINEMPIR= FREQ /NSAMP1;
DATA BOTH; MERGE TEMP FITNORM; BY I;*MERGE EMPIR & THEO DISTRIBUTIONS;
IF BINEMPIR=. THEN BINEMPIR=0;
IF NORMP LT .0001 AND BINEMPIR=0 THEN DELETE;
* NOW DATA SET SHOULD BE READY TO BE GRAPHED. VARS ARE I, BINEMPIR AND NORMP;
PROC PLOT; PLOT BINEMPIR*I='E' NORMP*I='T'/OVERLAY;
```

```
* LISTING OF SAS STATEMENTS FOR EXAMPLE 2;
DATA TDIST(KEEP=TEMPIR TDIS NSAMP N) THEO(KEEP=TD CUMTH TTHEO);
INPUT N NSAMP;
DO I=1 TO 15;TD=(I-8)*.4; *GENERATE PROBS FOR THEO T DIST;
CUMTH=PROBT(TD+.2,N-1);TTHEO=DIF1(CUMTH);OUTPUT THEO;END;
DO I=1 TO NSAMP;*GENERATE NSAMP EMPIRICAL T VALUES;
SUM=0; SS=0;
DO J=1 TO N;
X=NORMAL(13579); SUM+X; SS+X*X;
END;
SS=(N*SS-SUM*SUM)/(N*(N-1));
TEMPIR=SUM/SQRT(N*SS);
TDIS=FLOOR((TEMPIR+.2)*2.5)/2.5; *INTERVALIZE T'S;
IF TDIS LT -2.8 THEN TDIS=-2.8;IF TDIS GT 2.8 THEN TDIS=2.3;
OUTPUT TDIST; END;
CARDS;
20 500
PROC SUMMARY DATA=TDIST; CLASS TDIS;VAR NSAMP N;
OUTPUT OUT=TEMP MEAN=NSAMP1 N1;*CREATE FREQ DIST;
DATA NEXT(RENAME=(TDIS=TD));SET TEMP;
IF TYPE =0 THEN DELETE;
DROP NSAMP1 N1 TYPE FREQ;
TFREQ= FREQ /NSAMP1;*CREATE RELATIVE FREQ DIST;
DATA TOPLOT;MERGE NEXT THEO;BY TD; *MERGE EMPIR & THEO DISTRIBUTIONS;
IF TD=-2.8 THEN TTHEO=CUMTH;IF TD=2.8 THEN TTHEO=TTHEO+1-CUMTH;
PROC PLOT; PLOT TFREQ*TD='E' TTHEO*TD='T'/OVERLAY;
```

*LISTING OF SAS STATEMENTS FOR EXAMPLE 3;

```
DATA FISH(KEEP=R ZR RHO N NSAMP)
  THEO(KEEP=ZR ZR MUZ STERZ CUM SUMCUM NORMP);
INPUT N RHO NSAMP; R2=SQRT(1-RHO**2);
DO I=1 TO NSAMP; *GENERATE NSAMP VALUES OF R;
  SX=0; SSX=0; SY=0; SSY=0; SCP=0;
  DO J=1 TO N;
    X=NORMAL(12345); Y=RHO*X+R2*NORMAL(12345);
    SX=X; SSX=X*X; SY=Y; SSY=Y*Y; SCP=X*Y; END;
  R=(N*SCP-SX*SY)/SQRT((N*SSX-SX*SX)*(N*SSY-SY*SY));
  ZR=.5*LOG((1+R)/(1-R)); OUTPUT FISH; END;
  STERZ=SQRT(1/(N-3)); MUZ=.5*LOG((1+RHO)/(1-RHO)); SUMCUM=0;
  DO I=3 TO 16; *CALC NORMAL PROBABILITIES;
    ZR=.1*(I-1); ZZR=(ZR+.05-MUZ)/STERZ; CUM=PROBNORM(ZZR);
    IF I LE 15 THEN DO; NORMP=CUM-SUMCUM; SUMCUM=NORMP; OUTPUT THEO; END;
  ELSE DO; NORMP=1-SUMCUM; OUTPUT THEO; END;
END; CARDS;
50 .7 500
PROC RECODE DATA=FISH FUZZ=.05; *RECODE ZR TO 0.1 UNIT INTERVALS;
RECODE ZR
  LO THRU 0=.0 .1=.1 .2=.2 .3=.3 .4=.4 .5=.5 .6=.6 .7=.7 .8=.8
  .9=.9 1=.1 1.1=.1 1.2=.1 1.3 THRU HI=1.3;
PROC RECODE DATA=FISH FUZZ=.025; *RECODE R TO 0.05 UNIT INTERVALS;
RECODE R LO THRU .2=.2 .25=.25 .3=.3 .35=.35 .4=.4 .45=.45 .5=.5
  .55=.55 .6=.6 .65=.65 .7=.7 .75=.75 .8=.8 .85=.85 .9=.9 .95=.95 1=.1;
PROC SUMMARY DATA=FISH; CLASS R; VAR NSAMP; OUTPUT OUT=RDIST MEAN=NS1;
DATA PLOT; SET RDIST; IF TYPE=0 THEN DELETE; DROP _TYPE_ NS1 _FREQ_;
RPROP=FREQ/NS1; *CREATE REL. FREQ DIST FOR R;
PROC PLOT; PLOT RPROP*R;
PROC SUMMARY DATA=FISH; CLASS ZR; VAR NSAMP; OUTPUT OUT=ZRDIST MEAN=NS1;
DATA GETZR; SET ZRDIST; IF TYPE=0 THEN DELETE; DROP _TYPE_ NS1 _FREQ_;
ZRPROP=FREQ/NS1; *CREATE REL. FREQ DIST FOR ZR;
DATA BOTH; MERGE GETZR THEO; BY ZR; *MERGE EMPIR & THEO DIST AND PLOT;
PROC PLOT; PLOT ZRPROP*ZR='E' NORMP*ZR='T'/OVERLAY;
```

* LISTING OF SAS STATEMENTS FOR EXAMPLE 4;

```
DATA STUD; INPUT NAME $ 1-16; *CREATE STUDENT FILE; CARDS;
JONES E S
SMITH F G
ANDERSON R
RYAN J P
DATA SIM(KEEP=X1-X20 XBAR SD MUH MUT SIGMA Z PZ T PT);
ARRAY X X1-X20;
INPUT NSAMP MUT SIGMA MUH;
DO I=1 TO NSAMP; *CREATE OUTPUT DATA FOR EACH STUDENT;
  SUM=0; SS=0;
  DO OVER X;
    X=FLOOR(NORMAL(12345)*SIGMA+MUT+.5); SUM=X; SS=X*X; END;
  XBAR=SUM/20; SD=SQRT((SS*20-SUM*SUM)/380);
  Z=(XBAR-MUH)/(SIGMA/SQRT(20)); T=(XBAR-MUH)/(SD/SQRT(20));
  PZ=1-PROBNORM(Z); PT=1-PROBT(T,19); OUTPUT; END; CARDS;
4 50 10 45
*MERGE STUDENT NAME AND DATA FILES AND OUTPUT;
DATA NULL; MERGE STUD SIM; FILE PRINT;
PUT PAGE=40 OUTPUT FOR STUDENT NAME $16.//
  @20 *THE FOLLOWING IS A SAMPLE FROM A NORMAL DISTRIBUTION WITH'
  'A MEAN= MUT AND SIGMA= SIGMA//
  @30 'THE HYPOTHESIZED MEAN WAS MUH//
  @20 (X1-X20)(5.)// @20 'THE SAMPLE MEAN AND STANDARD DEVIATION'
  'ARE AS FOLLOWS: XBAR= 6.2 SD= 7.3//
  @20 'THE Z STATISTIC (KNOWN SIGMA) AND IT'S 1-TAILED P-VALUE'
```

' ARE: ' Z= 6.2 PZ= 7.3//
 @20 'THE T STATISTIC AND IT'S 1-TAILED P VALUE ARE: '
 T= 6.2 PT= 7.3;

* LISTING OF SAS STATEMENTS (AND DATA) FOR EXAMPLE 5;
 DATA ONE;ARRAY X X1-X10;ARRAY XD XD1-XD10;ARRAY KEY KEY1-KEY10;
 INPUT ID \$ 1-3 (X1-X10)(1.);RETAIN KEY1-KEY10;DROP KEY1-KEY10 X1-X10;
 IF N =1 THEN DO;*CREATE TEST KEY;
 DO OVER X;KEY=X;END;DELETE;END;
 ELSE DO;*SCORE ITEMS AND CREATE RAW SCORE FOR TEST;
 DO OVER X;XD=0;IF X=KEY THEN XD=1;END;
 TOT=SUM OF XD1-XD10;STAN=TOT;END;
 CARDS;
 KEY1234512345
 AAA5432154321
 BBB1111122222
 CCC1234554321
 DDD1232112321
 EEE1234512345
 FFF5434512345
 GGG1234544445
 HHH5534544445
 III1235421222
 JJJ1111121222
 PROC STANDARD MEAN=50 STD=10 OUT=TWO;VAR STAN;*CREATE STANDARD SCORES;
 PROC RANK TIES=MEAN GROUPS=100 OUT=THREE;VAR TOT; RANKS PCTLRK;
 * CREATE (APPROXIMATE) PERCENTILE RANK SCORES;
 PROC PRINT;VAR ID TOT STAN PCTLRK; *PRINT SCORE REPORT;
 PROC CORR ;VAR XD1-XD10;WITH TOT;*ITEM AND TEST STATISTICS;
 PROC MEANS NOPRINT;VAR XD1-XD10 TOT;*PREPARE FOR KR-20;
 OUTPUT OUT=FOUR VAR=V1-V10 VTOT;
 DATA FIVE;SET FOUR;SUMVAR=SUM(OF V1-V10);KR20=(10/9)*(1-SUMVAR/VTOT);
 PROC PRINT;VAR VTOT SUMVAR KR20; *OUTPUT KR-20;

*LISTING OF SAS STATEMENTS (AND DATA) FOR EXAMPLE 6;
 DATA PRE;ARRAY X X1-X5;ARRAY XD XD1-XD5;ARRAY KEY KEY1-KEY5;
 INPUT ID \$ 1-3 (X1-X5)(1.);RETAIN KEY1-KEY5;DROP KEY1-KEY5 X1-X5;
 IF N =1 THEN DO;DO OVER X;KEY=X;END;DELETE;END;*SET KEY FOR PRETEST;
 ELSE DO;
 DO OVER X;XD=0;IF X=KEY THEN XD=1;END;END;*SCORE PRETEST;
 CARDS;
 KEY12345
 AAA33243
 BBB22222
 CCC31422
 DDD52344
 EEE13321
 FFF22234
 GGG21115
 HHH12344
 III12222
 JJJ33222
 DATA POST;ARRAY Y Y1-Y5;ARRAY YD YD1-YD5;ARRAY KEY KEY1-KEY5;
 INPUT ID \$ 1-3 (Y1-Y5)(1.);RETAIN KEY1-KEY5;DROP KEY1-KEY5 Y1-Y5;
 IF N =1 THEN DO;DO OVER Y;KEY=Y;END;DELETE;END;*SET POSTTEST KEY;
 ELSE DO;
 DO OVER Y;YD=0;IF Y=KEY THEN YD=1;END;END;*SCORE POSTTEST;
 CARDS;
 KEY12345
 AAA13344

```

33812335
CCC22444
DDD12335
EEE12345
FFF32245
GGG12345
HHH13335
III13335
JJJ13324
DATA BOTH;MERGE PRE POST;BY ID;*MERGE PRETEST AND POSTTEST;
ARRAY YD YD1-YD5;ARRAY XD XD1-XD5;ARRAY A A1-A5;ARRAY D D1-D5;
DO OVER A;*CREATE INDICATORS OF PRE-POST CHANGE;
A=0;IF XD=0 AND YD=1 THEN A=1;
D=0;IF XD=1 AND YD=0 THEN D=1;END;
* CALC FREQ OF POSITIVE AND NEGATIVE CHANGE;
PROC MEANS NOPRINT;VAR A1-A5 D1-D5;OUTPUT OUT=TEMP SUM=R1-R5 W1-W5;
DATA MCNEM;SET TEMP;ARRAY R R1-R5;ARRAY W W1-W5;ARRAY MC MC1-MC5;
DO OVER R;MC=(R-W)/SQRT(R+W);END;
PROC PRINT;*OUTPUT CHANGE FREQUENCIES AND MCNEMAR'S Z STATISTIC;

* LISTING OF SAS STATEMENTS FOR EXAMPLE 7;
*READ DATA FOR 3 LEVELS OF REPEATED MEASURES FACTOR;
DATA RM;INPUT Y1-Y3;CARDS;
10 24 17
10 25 20
13 27 20
35 29 21
30 29 22
5 24 17
25 26 20
20 25 20
35 28 21
40 29 21
*NOW TO ANALYZE THE ABOVE MATRIX;PROC MATRIX;FETCH Y;
D=.7071 .40825/- .7071 .40825/0 -.8165;
NOBS=NROW(Y);NVAR=NCOL(Y);PRINT NOBS;PRINT NVAR; *GET PARAMETERS;
NV1=NVAR-1;NOBS1=NOBS-1;DFE=NV1#NOBS1; *MORE HOUSEKEEPING;
AVE=Y(+,)/NOBS;PRINT AVE; *CALC AND PRINT MEAN VECTOR;
COV=(Y'*Y-AVE*AVE#NOBS)/NOBS1;PRINT COV;*CALC COVARIANCE MATRIX;
AVED=AVE*D;COVD=D'*COV*D; *CALC MEAN AND COV MATRIX IN REDUCED SPACE;
NUM=TRACE(COVD)/2;DEN=SSQ(COVD)/NV1;EPSHAT=NUM/DEN; *CALC EPSHAT;
EPSTIL=(NOBS#NV1#EPSHAT-2)/(NV1*(NOBS1-NV1#EPSHAT)); *CALC EPSTIL;
MAUCHLW=DET(COVD)/(TRACE(COVD#NVAR))#NVAR; *CALC MAUCHLY SPHERICITY W;
* NOW TO OBTAIN THE UNIVARIATE SOLN;
MSTRT=(SSQ(Y(+,)))/NOBS-SUM(Y)2/(NOBS#NVAR)); *CALC SSTRT;
MSRES=(SSQ(Y)-SSQ(Y(+,)))/NVAR-MSTRT)/DFE; *CALC ERROR MS;
MSTRT=MSTRT/NV1; *CALC MSTRT;
FUV=MSTRT/MSRES; * UNIVARIATE F STATISTIC;PRINT FUV;PRINT NV1;PRINT DFE;
PROBU=1-PROBF(FUV,NV1,DFE);*UNIVARIATE P VALUE;PRINT PROBU;PRINT MAUCHLW;
DFN=NV1#EPSHAT;DFD=DFE#EPSHAT; *ADJUSTED DF FOR EPSHAT ADJ APPROX;
PROBESH=1-PROBF(FUV,DFN,DFD); *EPSHAT ADJ P VALUE;
PRINT EPSHAT;PRINT PROBESH;PRINT DFN;PRINT DFD;
DFN=NV1#EPSTIL;DFD=DFE#EPSTIL; *ADJUSTED DF FOR EPSTIL ADJ APPROX;
PROBEPST=1-PROBF(FUV,DFN,DFD); *EPSTIL ADJ P VALUE;
PRINT EPSTIL;PRINT PROBEPST;PRINT DFD;PRINT DFN;
* NOW FOR MULTIVARIATE SOLN;
TSQR=AVED*INV(COVD)*AVED#NOBS; * CALC T SQUARE STATISTIC;
FIV=TSQR/(NOBS-NV1)/(NOBS1#NV1); *F TRANSFORMATION OF T SQUARE;
PROBNV=1-PROBF(FIV,NV1,NOBS-NV1); *TV SOLN P VALUE;
PRINT TSQR;PRINT FIV;PRINT PROBNV;

```